Toronto, Canada

A Report on

Executive Summary of Module 2

Introduction to Data Analytics (ALY6000)

Guided by:

Prof. Mohammad Shafiqul Islam

Submitted by:

Name : Pratikkumar Indravadan Malaviya

NUID : 002963548

Date of submission : 28th January'2022

INTRODUCTION

The given module defines the concept of data visualization in R language. The phenomenon “Data Visualization” itself tells about the graphical representation of data. Although, it elaborates the key information towards the data in graphical format, likewise in different graphs. Moreover, the given dataset is all about the two Rocky Mountain lakes found in Alberta, which contain otoliths. "Usually, they are biomineralized ear stones that contribute to both hearing and vestibular function in fish". (*www.sciencedirect.com/topics/agricultural-and-biological-sciences/otoliths*). In addition, the data also includes age, fork length, lake information, and specific time duration in the era. As a result, there are certain graphs are plotted below with appropriate data specifications.

METHODOLOGY

The given dataset holds the time duration of otoliths in the era of 1977-80's where age and fork length become key attributes of fish. Therefore, in order to access this dataset, some packages are required. For example, the "FSA" (Fisheries stock assessment methods and data) and "FSAdata" to support the FSA package. This dataset is already present in R and it will automatically be imported while we install the aforementioned packages. The mentioned dataset is known as "BullTroutRML2". However, these findings are not enough to plot specified graphs, thus there are determined packages are also mandatory to install prior to illustration the graphs.

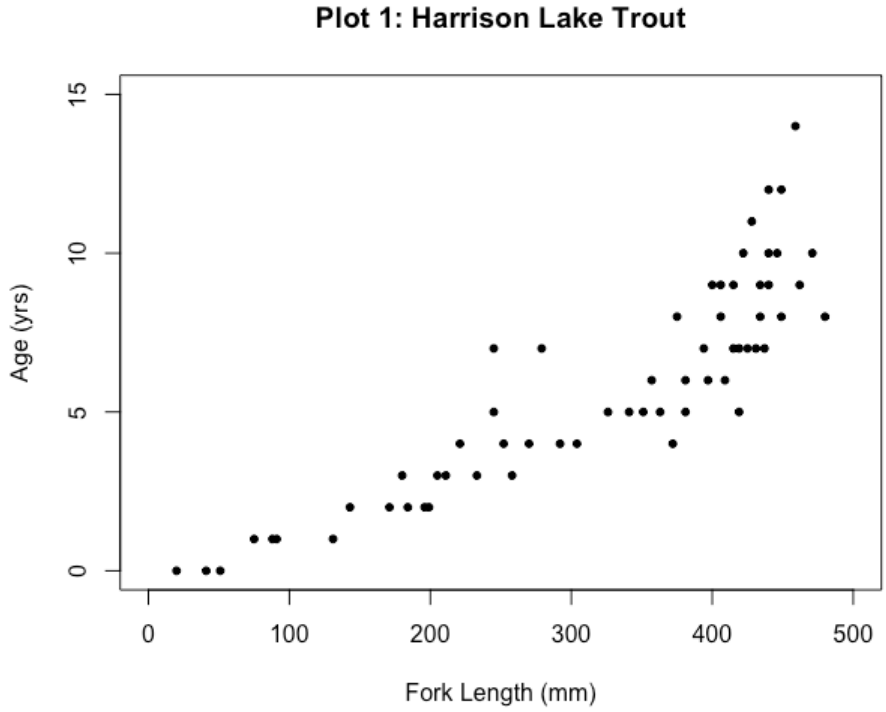
1. **Provide an analysis of descriptive characteristics of the data set provided by your instructor. This includes pertinent statistics including mean, median, quartiles, variance, standard deviation, skew, kurtosis, outliers etc. Include R console screen snipet to support your observations and conclusions. Below is a sample excerpt of an analysis of Harrison Lake fish from the BullTroutRML2 dataset.**

- After calculating basic statistics on BullTroutRML2, it denotes some values as an result.

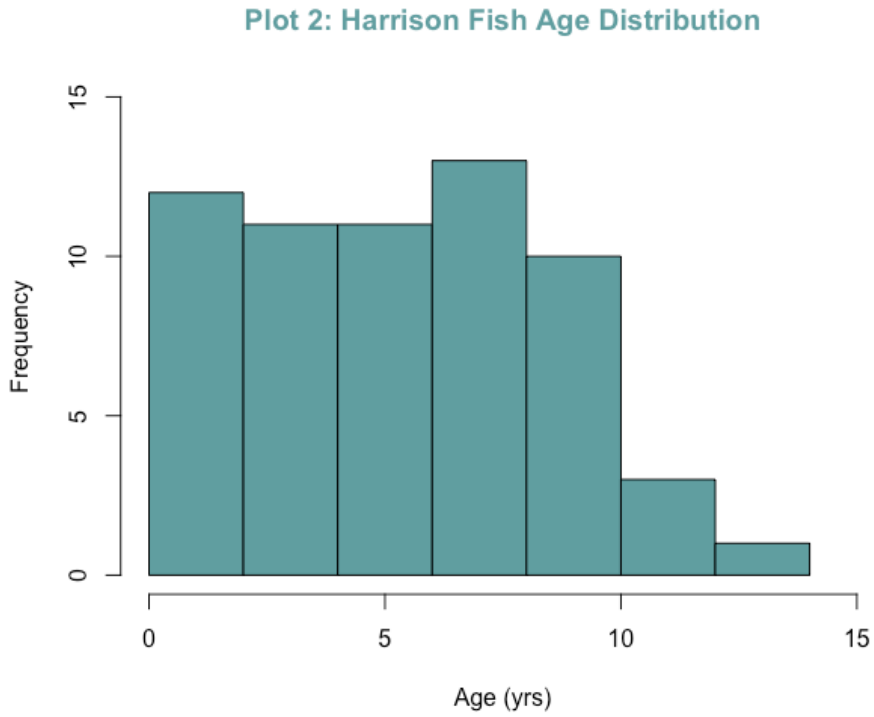
```
age               fl             lake         era
Min.   : 0.000   Min.   : 20   Harrison:61   1977-80:23
1st Qu.: 3.000   1st Qu.:221   Osprey  : 0   1997-01:38
Median : 6.000   Median :372
Mean   : 5.754   Mean   :319
3rd Qu.: 8.000   3rd Qu.:425
Max.   :14.000   Max.   :480
```

- The skewness and kurtosis are the measurement of the data which are used to describe shape of the data distribution. To calculate skewness we need to install 'moments' package, therefor we can directly calculate the skewness and kurtosis by executing its functions. In trout dataset we can calculate skew and kurtosis for otolith's age and its fork length.
- The outliers are basically denote a distance values form one observation to another observation. An outliers could be detect in descriptive statistics which contains maximum, minimum, histogram and so on.
- "Variance is the sum of the squares of differences between all number and means." *(https://www.geeksforgeeks.org/r-tutorial/?ref=lbp)*. We can calculate the variance in R by applying "var()" function.
- The standard deviation stands for square root of the variance. It measure the data values who's varies form the mean. Calculation of standard deviation in R take place with "sd()" function.

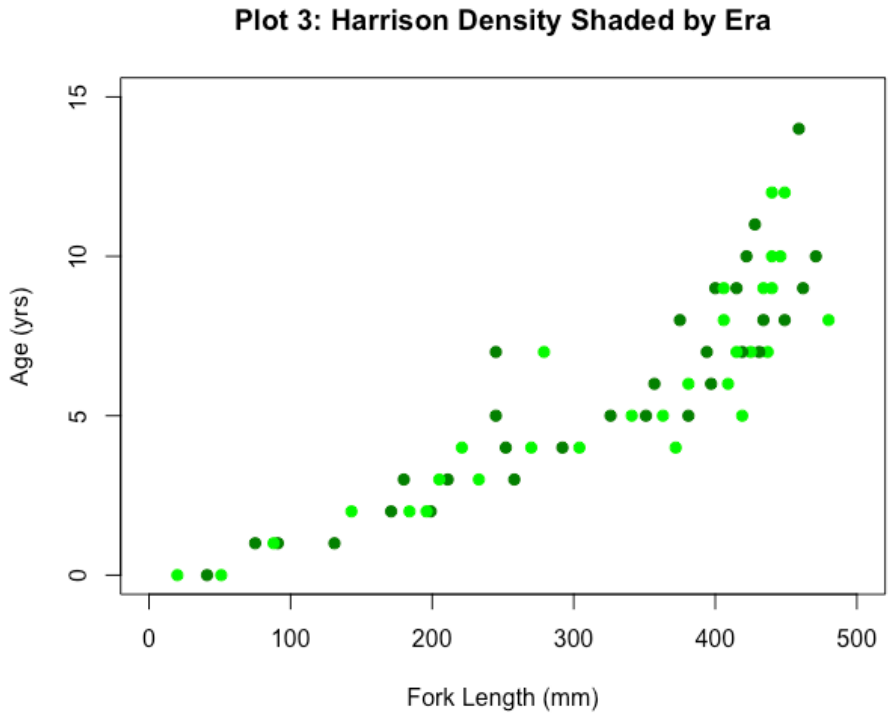
2. **Provide the executive with visualizations (at least 6) in that help them see the key characteristics you want to highlight. They can be boxplots, histograms, frequency and probability distributions, barplots (bar charts) or pareto. Not only is the goal to present your visual results, but also to explain the significance of what the visuals are displaying.**



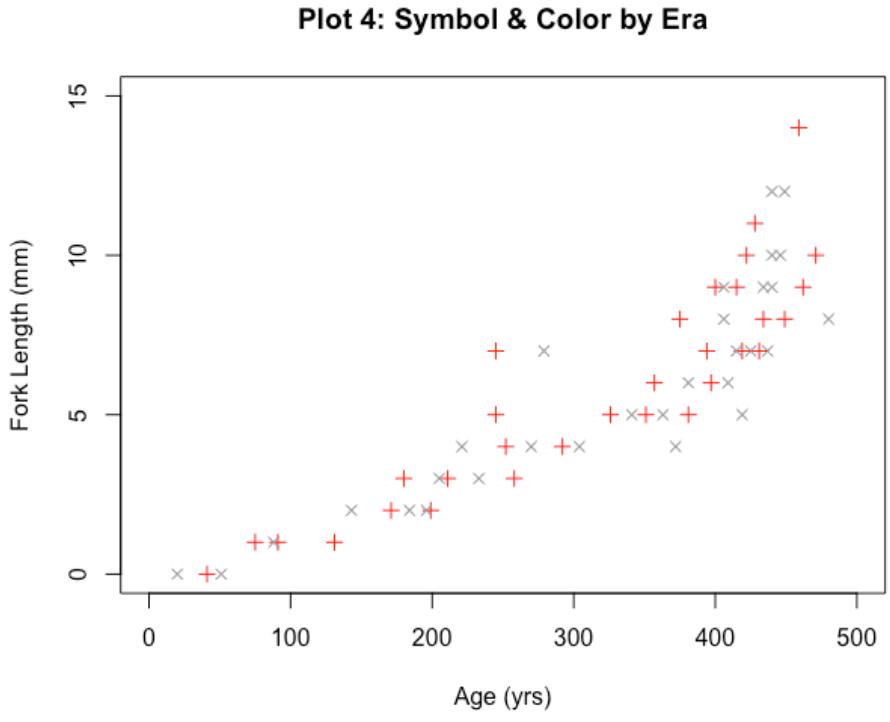
In the given Plot-1 (Scatter Plot) , we can say that the trout fish does contains fork length between 50-500(mm), on the other hand their ages are between 0-15(years). Thus, it is clear that the trout fishes can be fished when their age is greater than 5 years old and probability goes higher to get big fishes when they grownups.



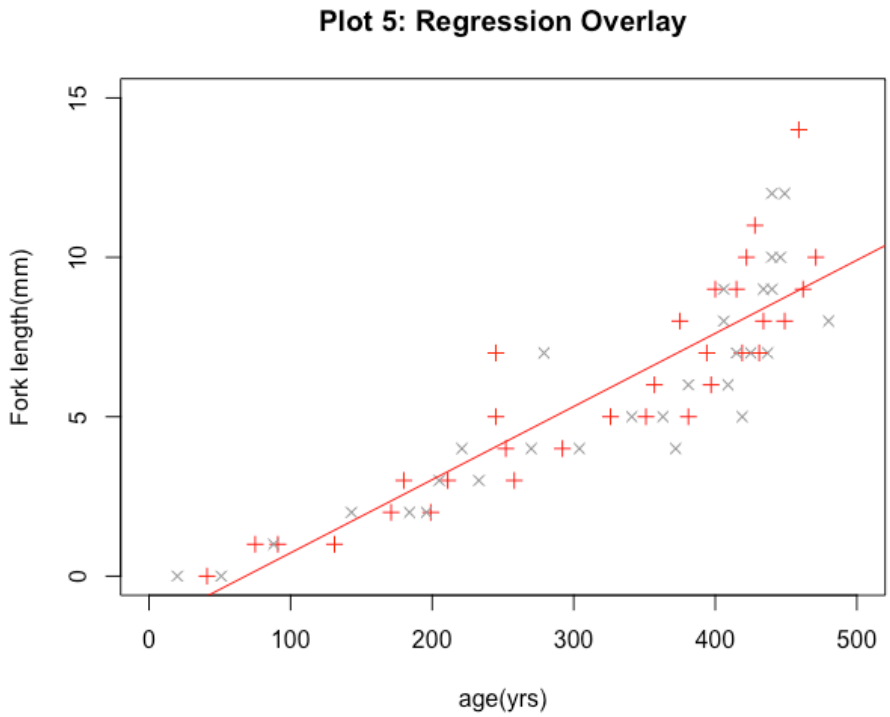
In given Plot-2 (Histogram Plot), the illustration indicates the age distribution of trout fish where it can be seen that fishes those who has ages less then 5-10 has higher frequency as compare to those fishes who are older then 10 years. The older fishes (10-15) has lower frequency ratio.



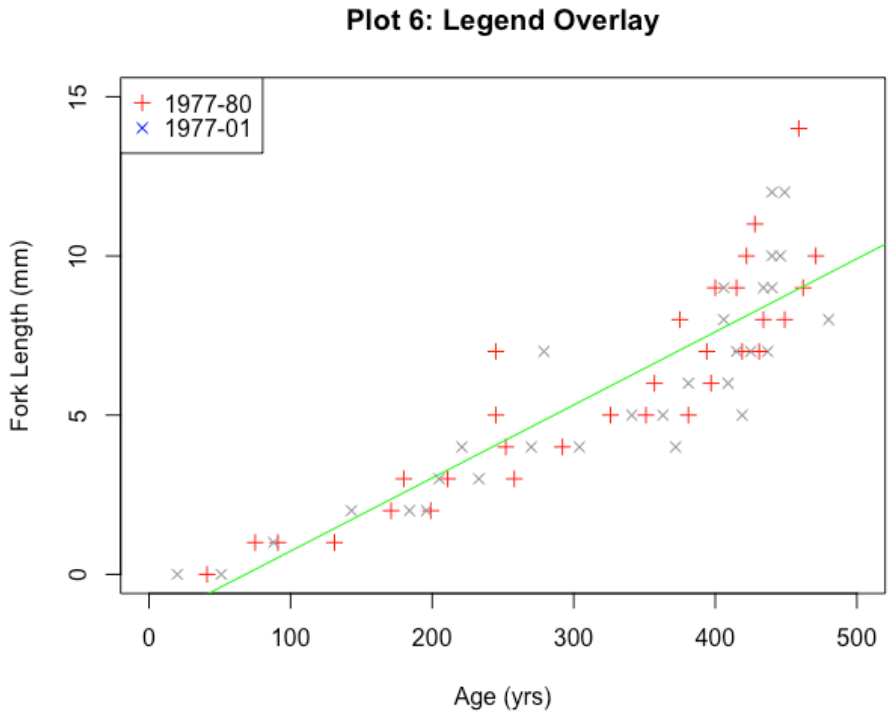
In the given Plot-3 (Overdense Plot), a compressed scatter plots determines clear visualization of age and fork length of the fishes with the era.



In the given Plot-4, the R language has pch symbol “+” and “x’ which are useful to assign some key information. These symbols does contain numeric values (3,4) where, we can out them as in argument. Moreover, the plot is highlighted with symbol colours to the era.



In the given Plot-5, we can clearly see that the regression line in red colour which describe the corresponding changes of y-variable on x-variable.



In the given Plot-6, we get to interact with the Legend, which are basically used to improvise the visuality of graph by identifying the lines, colours and symbols.

3. **Finally, provide a clear two to three sentence paragraph summary of the key points that you want the audience to walk away with regarding your analysis. This summary should present accurate analysis and be supported by the data presented in the rest of the report.**

- Although, the data set is also plays pivotal role in any analysis where sometimes it becomes easier to get work with data when we have sufficient amount of knowledge towards the data.
- Secondly, the stated analysis describes the importance of data visualization, where it defines key-points clearly in single visual instead of elaborating data theoretically.

- Moreover, the determined module express importance of statistics in data. The R language supports several predefined functions which can calculates the basic and complex statistics by simple execution of a function. Likewise mean, median, mode, quartiles, variance etc.
- The data statistics becomes supportive asset when we working with large datasets, based on these calculations the illustration of graph gets easier.

CONCLUSION

In this module, we understood the basics of statistics by implementing some normal operations on the given BullTroutRML2 dataset. In addition, we covered frequency distribution and data description by acknowledging common notations and their respective formulas. Furthermore, we learned how data visualization and its appropriate presentation is important to define the data in an alternative way. In R, we studied several plots likewise scatter plots, histogram plots, density plots, and many more. By providing effects for better visuality, we used different pch symbols, regression lines, and legends. Hence, the second module helps us to gain knowledge about how do these plots deliver us the key information of the given dataset.

BIBLIOGRAPHY

1. “Dplyr Tutorial | How to Filter Data Using Filter Function | R Programming Tutorial.” *YouTube*, YouTube, 2 June 2017, https://www.youtube.com/watch?v=yrVhA8GXvrc.
2. “How to Make a Scatterplot in R (with Regression Line).” *YouTube*, YouTube, 23 June 2017, https://www.youtube.com/watch?v=cGb5iqhf0NU.
3. “Dplyr Tutorial | How to Filter Data Using Filter Function | R Programming Tutorial.” *YouTube*, YouTube, 2 June 2017, https://www.youtube.com/watch?v=yrVhA8GXvrc.
4. Kabacoff, Robert. R In Action: Data Analysis and Graphics with R. Manning, 2015.
5. Tutorialspoint.com. 2022. *R Tutorial*. ***https://www.tutorialspoint.com/r/index.htm***
6. Skewness And Kurtosis In R Programming - GeeksforGeeks. (2020, April 29). GeeksforGeeks. https://www.geeksforgeeks.org/skewness-and-kurtosis-in-r-programming/.
7. Outliers Detection In R. (2020, August 11). R-bloggers. https://www.r-bloggers.com/2020/08/outliers-detection-in-r/.
8. Calculate the Average, Variance And Standard Deviation In R Programming - GeeksforGeeks. (2020, August 10). GeeksforGeeks. https://www.geeksforgeeks.org/calculate-the-average-variance-and-standard-deviation-in-r-programming/?ref=lbp.

APPENDIX

```
# 01 Name

print("Plotting Basics: Malaviya")


# 02 Install Packages

install.packages("FSA")

library(FSA)

install.packages("FSAdata")

library(FSAdata)

install.packages("magrittr")

library(magrittr)

install.packages("dplyr")

library(dplyr)

install.packages("plotrix")

library(plotrix)

install.packages("ggpubr")

library(ggpubr)


# standard deviation, skew, kurtosis, outliers

install.packages("moments")

library(moments)

data("BullTroutRML2")

skewness(BullTroutRML2$fl)

skewness(BullTroutRML2$age)
```

```
kurtosis(BullTroutRML2$fl)

kurtosis(BullTroutRML2$age)


var(BullTroutRML2$age) #Variance

var(BullTroutRML2$fl)


sd(BullTroutRML2$age) #Standard Deviation

sd(BullTroutRML2$fl)


# 03 Load the file

data("BullTroutRML2")


# 04 Display the 3 records

print(head(BullTroutRML2, n=3))


# 05 Removing specifies records

lake_harrison = filter(BullTroutRML2, lake == 'Harrison') # Filter Applied


# 06 Displaying first & last 5 records

 fisrt_five = head((lake_harrison),n=5)

 print(fisrt_five)

 last_five = tail((lake_harrison),n=5)

 print(last_five)


# 07-08 Structure & Summery
```

```
 print(data.frame(lake_harrison))
 print(summary(lake_harrison))


# 09 Scatter Plot
 plot(x = lake_harrison$fl,
      y = lake_harrison$age,
      xlab = 'Fork Length (mm)',
      ylab = 'Age (yrs)',
      xlim = c(0,500),
      ylim = c(0,15),
      pch = 20,
      main = "Plot 1: Harrison Lake Trout" )
 par(new=TRUE)


# 10 “Age” Histogram
 hist(x = lake_harrison$age,
      xlab = 'Age (yrs)',
      ylab = 'Freqency',
      xlim = c(0,15),
      ylim = c(0,15),
      col = 'cadetblue',
      main = 'Plot 2: Harrison Fish Age Distribution',
      col.main = 'cadetblue' )
 par(new=TRUE)
```

```
# 11 An Overdense Plot
 plot(x = lake_harrison$fl,
      y = lake_harrison$age,
      xlab = 'Fork Length (mm)',
      ylab = 'Age (yrs)',
      xlim = c(0,500),
      ylim = c(0,15),
      main = 'Plot 3: Harrison Density Shaded by Era',
      pch = 19,
      col = rgb(0,(1:2)/2,0),
      )

 par(new=TRUE)

# 12 Creating new object "tmp"
 tmp = headtail((lake_harrison),n=3)
 print(tmp)

# 13 Displaying the “era” column (variable) in the new “tmp” object
 print(tmp$era)

# 14  pchs vector
 pchs = c(3,4)
 pchs
```

```
# 15 cols vector
 cols = c("red","grey60")

# 16 Converting tmp into numeric
 class(tmp$era)
 as.numeric(tmp$era)

# 17 Initialization
 cols[tmp$era]

# 18 Plot 4: Symbol & Color by Era
 plot(age ~ fl,
      data = lake_harrison,
      xlim = c(0,500),
      ylim = c(0,15),
      main = "Plot 4: Symbol & Color by Era",
      xlab = "Age (yrs)",
      ylab = "Fork Length (mm)",
      pch = pchs,
      col = cols)
 par(new=TRUE)

 # 19 Plot 5: Regression Overlay".
 plot(age~fl,
      data = lake_harrison,
```

```
     main = 'Plot 5: Regression Overlay',
     xlab = "age(yrs)",
     ylab = "Fork length(mm)",
     xlim = c(0,500),
     ylim = c(0,15),
     pch = pchs,
     col = cols
     )
 abline(lm(lake_harrison$age~lake_harrison$fl),
       data=lake_harrison,col='red')

#20 Plot 6: Legend Overlay

 plot(age~fl,
     data = lake_harrison,
     main = 'Plot 6: Legend Overlay',
     xlim = c(0,500),
     ylim = c(0,15),
     xlab = "Age (yrs)",
     ylab = "Fork Length (mm)",
     pch = pchs,
     col = cols)

 abline(lm(lake_harrison$age ~ lake_harrison$fl),
       data=lake_harrison,col = 'green')
```

```
legend("topleft",
       legend = c('1977-80','1977-01'),
       col = c('red','blue'),
       pch = c(3,4)
       )
```

GITHUB

Username : pratik4511

Repository : https://github.com/pratik4511/malaviya_m2_project-2